Journal of Classification 32:85-113 (2015)
DOI: 10.1007/s00357-015-9175-1

The Generalized Linear Mixed Cluster-Weighted Model

Salvatore Ingrassia

University of Catania, Italy

Antonio Punzo

University of Catania, Italy

Giorgio Vittadini

University of Milano-Bicocca, Italy

Simona C. Minotti

University of Milano-Bicocca, Italy

Abstract: Cluster-weighted models (CWMs) are a flexible family of mixture models for fitting the joint distribution of a random vector composed of a response variable and a set of covariates. CWMs act as a convex combination of the products of the marginal distribution of the covariates and the conditional distribution of the response given the covariates. In this paper, we introduce a broad family of CWMs in which the component conditional distributions are assumed to belong to the exponential family and the covariates are allowed to be of mixed-type. Under the assumption of Gaussian covariates, sufficient conditions for model identifiability are provided. Moreover, maximum likelihood parameter estimates are derived using the EM algorithm. Parameter recovery, classification assessment, and performance of some information criteria are investigated through a broad simulation design. An application to real data is finally presented, with the proposed model outperforming other well-established mixture-based approaches.

Keywords: Cluster-weighted models; Model-based clustering, Generalized linear models, Mixed-type data.

The authors are grateful to the Editor and the reviewers for comments and suggestions which contributed to improve significantly the quality of the paper.

Authors' Addresses: S. Ingrassia and A. Punzo, Department of Economics and Business, University of Catania, Corso Italia, 55, 95129 Catania, Italy, email: s.ingrassia@unict.it, antonio.punzo@unict.it; G. Vittadini and S.C. Minotti, Department of Statistics, University of Milano-Bicocca, Via Bicocca degli Arcimboldi, 8, 20126 Milano, Italy, email: giorgio.vittadini@unimib.it; simona.minotti@unimib.it.

Published online: 18 April 2015

1. Introduction

Mixture models have attracted the increasing attention of numerous researchers in the recent decades (for a survey see, e.g., Titterington, Smith and Makov 1985, McLachlan and Peel 2000, Frühwirth-Schnatter 2006). In this paper, we focus on a family of mixture models, called cluster-weighted models (CWMs), proposed in a context of media technology under Gaussian assumptions (Gershenfeld 1997, 1999; Gershenfeld, Schöner and Metois 1999; Schöner 2000; Schöner and Gershenfeld 2001). CWMs are called saturated mixture regression models in Wedel (2002).

Let $(\boldsymbol{X}', Y)'$ be the pair of a vector of covariates $\boldsymbol{X}$ and a response variable Y defined on some space Ω with values in $\mathcal{X} \times \mathcal{Y}$. Assume that Ω can be partitioned into G groups, say $\Omega_1, \ldots, \Omega_G$. The CWM models the joint distribution $p(\boldsymbol{x}, y)$ of $(\boldsymbol{X}', Y)'$ as a convex combination, with weights $\pi_1, \ldots, \pi_G$, of the conditional distributions $p(y|\boldsymbol{x}, \Omega_g)$ times the marginal distributions $p(\boldsymbol{x}|\Omega_g)$. Expressed as a formula,

$$p(\boldsymbol{x}, y) = \sum_{g=1}^{G} p(y|\boldsymbol{x}, \Omega_g)\, p(\boldsymbol{x}|\Omega_g)\, \pi_g. \tag{1}$$

Quite recently, Ingrassia, Minotti and Vittandini (2012) reformulated the CWM in a statistical setting and showed that it is a general and flexible family of mixture models. In particular, they prove that, under suitable assumptions, if both $p(y|\boldsymbol{x}, \Omega_g)$ and $p(\boldsymbol{x}|\Omega_g)$ are Gaussian, then mixtures of Gaussian distributions on $(\boldsymbol{X}', Y)'$, mixtures of linear Gaussian regressions, and mixtures of linear Gaussian regressions with concomitant variables, using $\boldsymbol{X}$ as the concomitant variable, can be considered as nested in the linear Gaussian CWM. Moreover, in the same paper, the linear t-CWM was introduced considering both $p(y|\boldsymbol{x}, \Omega_g)$ and $p(\boldsymbol{x}|\Omega_g)$ to be t-distributed; again, mixture of t distributions and mixtures of regression models with t errors can be considered as nested in the linear t-CWM. Subsequently, Ingrassia, Minotti and Punzo (2014) presented a family of twelve CWMs, nested in the linear t-CWM, for model-based clustering. Subedi, Punzo, Ingrassia, and McNicholas (2013) addressed the problem of applicability of the CWM in high-dimensional $\boldsymbol{X}$-spaces by assuming latent factors for the covariates in each mixture component. Finally, to allow the model to also be applied to groups having nonlinear dependencies of Y on $\boldsymbol{x}$, Punzo (2014) proposed the polynomial Gaussian CWM.

However, in many practical cases, we are faced with categorical or discrete responses. For example, in healthcare studies, a typical response is the length of stay, while in the educational framework, the response is often the item-category chosen in some test. In the literature about mixture mod-

els, such problems are usually approached considering mixtures of generalized linear models (see, e.g., McLachlan 1997, McLachlan and Peel 2000, Wedel and De Sarbo 1995). We remark that Gershenfeld (1999) also coped with the problem of discrete sets of values, such as events, patterns, or conditions, but without really modeling the joint probability of the dependent variable and the covariates. Moreover, we are often faced with covariates of mixed-type (continuous and finite discrete).

The rest of the paper is organized as follows. In Section 2, we introduce a broad family of CWMs where the component conditional distributions are assumed to belong to the exponential family, and where the covariates are allowed to be of mixed-type (by using a Gaussian distribution for the continuous covariates and the product of multinomial distributions for the finite discrete covariates). In Section 3, we give sufficient conditions for CWMs, with continuous covariates only, to be identifiable. In Section 4, we describe an expectation-maximization (EM) algorithm for maximum likelihood parameter estimation and, in Section 5, we analyze parameter recovery via a broad simulation study. Likelihood-based information criteria can be adopted to select the number of mixture components G, and the performance of some of them is investigated in Section 6. A real data set is analyzed in Section 7 and some concluding remarks are presented in Section 8.

2. The Model

Suppose that the vector of covariates can be written as $\boldsymbol{X} = (\boldsymbol{U}', \boldsymbol{V}')'$, where $\boldsymbol{U}$ is a p-variate vector of continuous covariates and $\boldsymbol{V}$ is a q-variate vector of finite discrete covariates, with values $c_1, \ldots, c_q$, respectively, being $p + q = d$. In this case, $\mathcal{X} = \mathbb{R}^p \times \{1, \ldots, c_1\} \times \cdots \times \{1, \ldots, c_q\}$. Naturally, special cases concern either $p = 0$ or $q = 0$. Moreover, with reference to (1) and based on the classical latent class model (LCM; see Vermunt and Magidson 2002 and Hennig and Liao 2013), assume that $\boldsymbol{U}$ and $\boldsymbol{V}$ are "locally" independent; that is, they are independent within each mixture component. Model (1) can be so written as

$$\begin{aligned} p(\boldsymbol{x}, y; \boldsymbol{\vartheta}) &= \sum_{g=1}^{G} q(y|\boldsymbol{x}; \boldsymbol{\xi}_g) p(\boldsymbol{x}; \boldsymbol{\psi}_g) \pi_g \\ &= \sum_{g=1}^{G} q(y|\boldsymbol{x}; \boldsymbol{\xi}_g) p\left(\boldsymbol{u}; \boldsymbol{\psi}_g^*\right) p\left(\boldsymbol{v}; \boldsymbol{\psi}_g^{**}\right) \pi_g, \end{aligned} \tag{2}$$

where $q(y|\boldsymbol{x}; \boldsymbol{\xi}_g)$ denotes the conditional density of $Y|\boldsymbol{x}, \Omega_g$ with parameter $\boldsymbol{\xi}_g$, $p(\boldsymbol{x}; \boldsymbol{\psi}_g)$ is the marginal distribution of $\boldsymbol{X}|\Omega_g$ with parameter $\boldsymbol{\psi}_g$, $p\left(\boldsymbol{u}; \boldsymbol{\psi}_g^*\right)$ is the marginal distribution of $\boldsymbol{U}$ with parameter $\boldsymbol{\psi}_g^*$, $p\left(\boldsymbol{v}; \boldsymbol{\psi}_g^{**}\right)$

is the marginal distribution of $\boldsymbol{V}$ with parameter $\boldsymbol{\psi}_g^{**}$, $g = 1, \ldots, G$, and $\boldsymbol{\vartheta}$ contains all of the parameters of the model.

2.1 Modelling for $q(y|\boldsymbol{x}; \boldsymbol{\xi}_g)$

In order to deal with various response types, we assume that $q\left(y|\boldsymbol{x}; \boldsymbol{\xi}_g\right)$ belongs to the exponential family. Thus, in general, $\mathcal{Y} \subseteq \mathbb{R}$. It is well known that the exponential family is strictly related to the generalized linear models (see McCullagh and Nelder 2000), where a monotone and differentiable link function $h(\cdot)$ is introduced to relate the expected value μ_g, of $Y|\Omega_g$, to the covariates $\boldsymbol{X}$ through the relation $h(\mu_g) = \beta_{0g} + \boldsymbol{\beta}_{1g}'\boldsymbol{x}$. Because the interest is now in the parameters $(\beta_{0g}, \boldsymbol{\beta}_{1g}')' = \boldsymbol{\beta}_g$, the distribution of $Y|\boldsymbol{x}, \Omega_g$ will be denoted by $q(y|\boldsymbol{x}; \boldsymbol{\beta}_g, \lambda_g)$, where λ_g is an additional parameter to take into account when a distribution from a two-parameter exponential family is considered.

Quite often, the interest is in modeling discrete responses. Therefore, the following focuses mainly on two particular members of the exponential family: the Poisson and the binomial.

The binomial CWM. Assume that Y takes values in $\mathcal{Y} = \{0, 1, \ldots, M\}$, for some given $M \in \mathbb{N}$, and that $Y|\boldsymbol{x}, \Omega_g$ is binomial with parameters $\left(M, \mu_g\left(\boldsymbol{x}; \boldsymbol{\beta}_g\right)/M\right)$; that is, $Y|\boldsymbol{x}, \Omega_g \sim \text{Bin}\left(M, \mu_g\left(\boldsymbol{x}; \boldsymbol{\beta}_g\right)/M\right)$. In this case,

$$q\left(y|\boldsymbol{x}; \boldsymbol{\beta}_g\right) = \binom{M}{y} \left[\frac{\mu_g\left(\boldsymbol{x}; \boldsymbol{\beta}_g\right)}{M}\right]^y \left[1 - \frac{\mu_g\left(\boldsymbol{x}; \boldsymbol{\beta}_g\right)}{M}\right]^{M-y}, \quad (3)$$

where

$$\mu_g\left(\boldsymbol{x}; \boldsymbol{\beta}_g\right) = M\frac{\exp(\beta_{0g} + \boldsymbol{\beta}_{1g}'\boldsymbol{x})}{1 + \exp(\beta_{0g} + \boldsymbol{\beta}_{1g}'\boldsymbol{x})}.$$

Model (2), with conditional distributions (3), will be called the binomial CWM. Note that, the repetition parameter M is the same for all mixture components.

The Poisson CWM. Assume that Y takes values in $\mathcal{Y} = \mathbb{N}$ and that $Y|\boldsymbol{x}, \Omega_g$ is Poisson with parameter $\mu_g\left(\boldsymbol{x}; \boldsymbol{\beta}_g\right)$; that is, $Y|\boldsymbol{x}, \Omega_g \sim \text{Poi}\left[\mu_g\left(\boldsymbol{x}; \boldsymbol{\beta}_g\right)\right]$. In this case,

$$q(y|\boldsymbol{x}; \boldsymbol{\beta}_g) = \exp\left[-\mu_g(\boldsymbol{x}; \boldsymbol{\beta}_g)\right] \frac{\left[\mu_g\left(\boldsymbol{x}; \boldsymbol{\beta}_g\right)\right]^y}{y!}, \quad (4)$$

where

$$\mu_g(\boldsymbol{x};\boldsymbol{\beta}_g) = \exp\left(\beta_{0g} + \boldsymbol{\beta}'_{1g}\boldsymbol{x}\right).$$

Model (2), with conditional distributions (4), will be named the Poisson CWM.

2.2 Modelling for $p(\boldsymbol{x};\boldsymbol{\psi}_g)$

The term $p\left(\boldsymbol{u};\boldsymbol{\psi}_g^*\right)$ in (2) is modeled here according to a p-variate Gaussian density with mean $\boldsymbol{\mu}_g$ and covariance matrix $\boldsymbol{\Sigma}_g$, i.e., $p\left(\boldsymbol{u};\boldsymbol{\psi}_g^*\right) = \phi\left(\boldsymbol{u};\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g\right)$. With respect to the term $p(\boldsymbol{v};\boldsymbol{\psi}_g^{**})$ in (2), assume that each finite discrete covariate in $\boldsymbol{V}$ can be represented by a binary vector $\boldsymbol{v}^r = \left(v^{r1},\ldots,v^{rc_r}\right)'$, where $v^{rs} = 1$ if v_r is equal to the value s, with $s \in \{1,\ldots,c_r\}$, and $v^{rs} = 0$ otherwise. Furthermore, assume that the q finite discrete covariates are independent given the mixture component. Then, we have

$$p(\boldsymbol{v};\boldsymbol{\alpha}_g) = \prod_{r=1}^{q}\prod_{s=1}^{c_r}\left(\alpha_{grs}\right)^{v^{rs}}, \quad g = 1,\ldots,G, \tag{5}$$

where $\boldsymbol{\alpha}_g = \left(\boldsymbol{\alpha}'_{g1},\ldots,\boldsymbol{\alpha}'_{gq}\right)'$, with $\boldsymbol{\alpha}_{gr} = \left(\alpha_{gr1},\ldots,\alpha_{grc_q}\right)'$, $\alpha_{grs} > 0$, and $\sum_{s=1}^{c_r}\alpha_{grs}$, $r = 1,\ldots,q$. In particular, the density $p\left(\boldsymbol{v};\boldsymbol{\alpha}_g\right)$ in (5) is given by the product of q conditionally independent multinomial distributions of parameters $\boldsymbol{\alpha}_{gr}$, $r = 1,\ldots,q$.

2.3 The Resulting Overall Model

Based on the above assumptions, model (2) assumes the form

$$p\left(\boldsymbol{x},y;\boldsymbol{\vartheta}\right) = \sum_{g=1}^{G} q\left(y|\boldsymbol{x};\boldsymbol{\beta}_g,\lambda_g\right)\phi\left(\boldsymbol{u};\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g\right)p\left(\boldsymbol{v};\boldsymbol{\alpha}_g\right)\pi_g. \tag{6}$$

It will be named the generalized linear mixed CWM hereafter; the prefix "generalized linear" refers to the local relation of Y given $\boldsymbol{x}$, while the term "mixed" underlines the mixed-type nature of the random covariates. The special case without finite discrete covariates, that is

$$p\left(\boldsymbol{u},y;\boldsymbol{\vartheta}\right) = \sum_{g=1}^{G} q(y|\boldsymbol{u};\boldsymbol{\beta}_g,\lambda_g)\phi(\boldsymbol{u};\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g)\pi_g, \tag{7}$$

will be hereafter referred to as generalized linear Gaussian CWM.

2.4 Related Mixture Models

Some of the existing mixture models with covariates are related to the generalized linear mixed CWM. Examples include *i)* mixtures of generalized linear models (see, e.g., Wedel and De Sarbo 1995, McLachlan 1997), having conditional distribution

$$q(y|\boldsymbol{x};\boldsymbol{\vartheta}) = \sum_{g=1}^{G} q(y|\boldsymbol{x};\boldsymbol{\beta}_g,\lambda_g)\pi_g, \tag{8}$$

where $\boldsymbol{\vartheta} = \{\boldsymbol{\beta}_g, \lambda_g, \pi_g; g = 1,\ldots,G\}$, and *ii)* mixtures of generalized linear models with concomitant variables (see, e.g., Grün and Leisch 2008b), which use the covariates as concomitant variables, having conditional distribution

$$q(y|\boldsymbol{x};\boldsymbol{\vartheta}) = \sum_{g=1}^{G} q(y|\boldsymbol{x};\boldsymbol{\beta}_g,\lambda_g)p\left(\Omega_g|\boldsymbol{x};\boldsymbol{w}\right), \tag{9}$$

where the component weight $p(\Omega_g|\boldsymbol{x};\boldsymbol{w})$ is now function of $\boldsymbol{x}$ through some parameter $\boldsymbol{w}$, and where $\boldsymbol{\vartheta} = \{\boldsymbol{\beta}_g, \lambda_g, \boldsymbol{w}_g; g = 1,\ldots,G\}$. The probability $p(\Omega_g|\boldsymbol{x};\boldsymbol{w})$ is usually modeled by a multinomial logistic distribution

$$p(\Omega_g|\boldsymbol{x};\boldsymbol{w}) = \frac{\exp(w_{0g} + \boldsymbol{w}'_{1g}\boldsymbol{x})}{\sum_{j=1}^{G}\exp(w_{0j} + \boldsymbol{w}'_{1j}\boldsymbol{x})},$$

where $\boldsymbol{w}_g = (w_{0g}, \boldsymbol{w}'_{1g})'$.

3. Identifiability

In order to estimate the parameters of model (6), it is important to establish its identifiability. General conditions for identifiability of mixtures of linear models can be found in Hennig (2000); based on such results, Grün and Leisch (2008a) provided results about identifiability for model (8). Follmann and Lambert (1991) and Wang (1994) established identifiability results for mixtures of logistic regression models (the latter paper considered the case with concomitant variables). Related results are also summarized in Frühwirth-Schnatter (2006).

Identifiability for mixture models can be defined as follows. Consider a parametric class of density (probability) functions $\mathcal{F} = \{f(\boldsymbol{z};\boldsymbol{\psi}) : \boldsymbol{z} \in \mathcal{Z}\,,$

$\boldsymbol{\psi} \in \boldsymbol{\Psi}\}$ and then the class of finite mixtures of functions in $\mathcal{F}$,

$$\mathcal{H} = \Bigg\{ h(\boldsymbol{z};\boldsymbol{\vartheta}) : h(\boldsymbol{z};\boldsymbol{\vartheta}) = \sum_{g=1}^{G} f(\boldsymbol{z};\boldsymbol{\psi}_g)\pi_g, \text{ with } \pi_g > 0 \text{ and } \sum_{g=1}^{G} \pi_g = 1,$$

$$f(\cdot;\boldsymbol{\psi}_g) \in \mathcal{F}, g = 1,\ldots,G, \boldsymbol{\psi}_g \neq \boldsymbol{\psi}_j \text{ for } g \neq j, G \in \mathbb{N}, \boldsymbol{z} \in \mathcal{Z}, \boldsymbol{\vartheta} \in \boldsymbol{\Theta} \Bigg\}.$$

This class is identifiable if, given two members

$$h(\boldsymbol{z};\boldsymbol{\vartheta}) = \sum_{g=1}^{G} f(\boldsymbol{z};\boldsymbol{\psi}_g)\pi_g \quad \text{and} \quad h(\boldsymbol{z};\widetilde{\boldsymbol{\vartheta}}) = \sum_{s=1}^{\widetilde{G}} f(\boldsymbol{z};\widetilde{\boldsymbol{\psi}}_s)\widetilde{\pi}_s$$

of $\mathcal{H}$, the equality $h(\boldsymbol{z};\boldsymbol{\vartheta}) = h(\boldsymbol{z};\widetilde{\boldsymbol{\vartheta}})$ implies that $G = \widetilde{G}$ and for each $g \in \{1,\ldots,G\}$ there exists $s \in \{1,\ldots,G\}$ such that $\pi_g = \widetilde{\pi}_s$ and $\boldsymbol{\psi}_g = \widetilde{\boldsymbol{\psi}}_s$.

Here, we face with the identifiability issue for the generalized linear Gaussian CWM defined in (7). In particular, we establish a sufficient condition for the identifiability of the class

$$C = \Bigg\{ p(\boldsymbol{u},y;\boldsymbol{\vartheta}) \,:\, p(\boldsymbol{u},y;\boldsymbol{\vartheta}) = \sum_{g=1}^{G} q(y|\boldsymbol{u};\boldsymbol{\beta}_g,\lambda_g)\phi(\boldsymbol{u};\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g)\pi_g,$$

$$\text{with } \pi_g > 0,$$

$$\sum_{g=1}^{G} \pi_g = 1,\ (\boldsymbol{\beta}_g,\lambda_g) \neq (\boldsymbol{\beta}_j,\lambda_j) \text{ for } g \neq j, \left(\boldsymbol{u}',y\right)' \in \mathbb{R}^p \times \mathcal{Y},$$

$$\boldsymbol{\vartheta} = \{\boldsymbol{\beta}_g,\lambda_g,\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g,\pi_g; g = 1,\ldots,G\} \in \boldsymbol{\Theta}, G \in \mathbb{N} \Bigg\}, \quad (10)$$

where $\mathcal{Y}$ depends on the component distribution q. In the following theorem, we provide sufficient conditions for $\mathcal{C}$ to be identifiable in $\mathcal{U} \times \mathcal{Y}$, where $\mathcal{U} \subseteq \mathbb{R}^p$ is a set having probability one according to the p-variate Gaussian density ϕ. In other words, we prove that the class $\mathcal{C}$ is identifiable for almost all $\boldsymbol{u} \in \mathbb{R}^p$ and for all $y \in \mathcal{Y}$.

Theorem 1. *Let $\mathcal{C}$ be the class defined in* (10) *and assume that there exists a set $\mathcal{U} \subseteq \mathbb{R}^p$ having probability one according to the p-variate Gaussian density such that the mixture of generalized linear models*

$$\sum_{g=1}^{G} q(y|\boldsymbol{u};\boldsymbol{\beta}_g,\lambda_g)\alpha_g(\boldsymbol{u}), \quad y \in \mathcal{Y}, \quad (11)$$

is identifiable for each fixed $\boldsymbol{u} \in \mathcal{U}$*, where* $\alpha_1(\boldsymbol{u}), \ldots, \alpha_G(\boldsymbol{u})$ *are positive weights summing to one for each* $\boldsymbol{u} \in \mathcal{U}$*. Then the class* $\mathcal{C}$ *is identifiable in* $\mathcal{U} \times \mathcal{Y}$*. Proof:* The proof is given in the Appendix.

Note that, when the distribution q is binomial, the repetition parameter M has to be checked because mixtures of binomials with the same repetition parameter M are identifiable if and only if $G \leq (M+1)/2$; see Teicher (1963).

4. The EM Algorithm for Parameter Estimation

Let $(\boldsymbol{x}_1', y_1)', \ldots, (\boldsymbol{x}_n', y_n)'$ be a sample of n independent observation pairs drawn from model (6). The corresponding likelihood, for a fixed number of components G, is given by

$$L(\boldsymbol{\vartheta}) = \prod_{i=1}^{n} p(\boldsymbol{x}_i, y_i; \boldsymbol{\vartheta}) = \prod_{i=1}^{n} \sum_{g=1}^{G} \pi_g q(y_i|\boldsymbol{x}_i; \boldsymbol{\beta}_g, \lambda_g) \phi(\boldsymbol{u}_i; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) p(\boldsymbol{v}_i; \boldsymbol{\alpha}_g).$$

Define $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iG})'$, with $z_{ig} = 1$ if $(\boldsymbol{x}_i', y_i)'$ comes from Ω_g, and $z_{ig} = 0$ otherwise, and consider the complete data $\{(\boldsymbol{x}_i', y_i, \boldsymbol{z}_i')'; i = 1, \ldots, n\}$. Then, the complete-data likelihood can be written as

$$L_c(\boldsymbol{\vartheta}) = \prod_{i=1}^{n} \prod_{g=1}^{G} \left[q(y_i|\boldsymbol{x}_i; \boldsymbol{\beta}_g, \lambda_g) \phi(\boldsymbol{u}_i; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) p(\boldsymbol{v}_i; \boldsymbol{\alpha}_g) \pi_g \right]^{z_{ig}}. \quad (12)$$

The corresponding complete-data log-likelihood, the logarithm of (12), can be written as

$$\begin{aligned} & l_c(\boldsymbol{\vartheta}) \\ &= \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \left[\ln q(y_i|\boldsymbol{x}_i; \boldsymbol{\beta}_g, \lambda_g) + \ln \phi(\boldsymbol{u}_i; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) + \ln p(\boldsymbol{v}_i; \boldsymbol{\alpha}_g) + \ln \pi_g \right] \\ &= \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \ln q(y_i|\boldsymbol{x}_i; \boldsymbol{\beta}_g, \lambda_g) + \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \ln \phi(\boldsymbol{u}_i; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) + \\ &\quad + \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \ln p(\boldsymbol{v}_i; \boldsymbol{\alpha}_g) + \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \ln \pi_g. \end{aligned} \quad (13)$$

Maximization of $L(\boldsymbol{\vartheta})$, through $L_c(\boldsymbol{\vartheta})$, is here achieved by the EM algorithm (Dempster, Laird and Rubin 1977). Each iteration of the EM algorithm alternates between two steps, the E-step (expectation) and the M-step (maximization), that will be described below for model (6).

4.1 E-step

The E-step, on the $(k+1)$th iteration, $k = 0, 1, \ldots$, requires calculation of the expectation of $l_c(\boldsymbol{\vartheta})$ given the observed data and the provisional estimate $\boldsymbol{\vartheta}^{(k)}$, of $\boldsymbol{\vartheta}$, arising from the previous iteration. As $l_c(\boldsymbol{\vartheta})$ is linear in the unobservable data z_{ig}, the E-step simply requires calculation of the current conditional expectation of Z_{ig} given the observed sample, where Z_{ig} is the random variable corresponding to z_{ig}. In particular, for $i = 1, \ldots, n$ and $g = 1, \ldots, G$, it follows that

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\vartheta}^{(k)}}\left[Z_{ig}\left|\left(\boldsymbol{x}_i', y_i\right)'\right.\right] &= \tau_{ig}^{(k)} \\ &= \frac{q\left(y_i|\boldsymbol{x}_i;\boldsymbol{\beta}_g^{(k)},\lambda_g^{(k)}\right)\phi\left(\boldsymbol{u}_i;\boldsymbol{\mu}_g^{(k)},\boldsymbol{\Sigma}_g^{(k)}\right)p\left(\boldsymbol{v}_i;\boldsymbol{\alpha}_g^{(k)}\right)\pi_g^{(k)}}{p\left(\boldsymbol{x}_i,y_i;\boldsymbol{\vartheta}^{(k)}\right)}, \quad (14)\end{aligned}$$

which corresponds to the posterior probability that the unlabeled observation $(\boldsymbol{x}_i', y_i)'$ belongs to the gth component of the mixture, using the current fit $\boldsymbol{\vartheta}^{(k)}$ for $\boldsymbol{\vartheta}$.

4.2 M-step

In the M-step, on the $(k+1)$th iteration, $k = 0, 1, \ldots$, the conditional expectation of $l_c(\boldsymbol{\vartheta})$ given the observed data, say $Q(\boldsymbol{\vartheta}; \boldsymbol{\vartheta}^{(k)})$, is maximized with respect to $\boldsymbol{\vartheta}$. To this end, the values z_{ig} in (13) are simply replaced by their current expectations $\tau_{ig}^{(k)}$ obtained in (14), yielding

$$\begin{aligned}Q\left(\boldsymbol{\vartheta};\boldsymbol{\vartheta}^{(k)}\right) = &\sum_{i=1}^{n}\sum_{g=1}^{G}\tau_{ig}^{(k)}\ln\pi_g + \sum_{i=1}^{n}\sum_{g=1}^{G}\tau_{ig}^{(k)}\ln q(y_i|\boldsymbol{x}_i;\boldsymbol{\beta}_g,\lambda_g) + \\ &+\sum_{i=1}^{n}\sum_{g=1}^{G}\tau_{ig}^{(k)}\ln\phi\left(\boldsymbol{u}_i;\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g\right) + \sum_{i=1}^{n}\sum_{g=1}^{G}\tau_{ig}^{(k)}\ln p\left(\boldsymbol{v}_i;\boldsymbol{\alpha}_g\right). \quad (15)\end{aligned}$$

As the four terms on the right-hand side have zero cross-derivatives, they can be maximized separately. Let us set $\boldsymbol{\pi} = \{\pi_g; g = 1, \ldots, G\}$, $\boldsymbol{\beta} = \{\boldsymbol{\beta}_g; g = 1, \ldots, G\}$, and $\boldsymbol{\lambda} = \{\lambda_g; g = 1, \ldots, G\}$.

4.2.1 Mixture Weights

The maximum of $Q(\boldsymbol{\vartheta}; \boldsymbol{\vartheta}^{(k)})$ with respect to $\boldsymbol{\pi}$, subject to the constraints on these parameters, is achieved by maximizing the augmented function

$$\sum_{i=1}^{n}\sum_{g=1}^{G}\tau_{ig}^{(k)}\ln\pi_g-\eta\left(\sum_{g=1}^{G}\pi_g-1\right), \tag{16}$$

where η is a Lagrangian multiplier. Setting the derivative of equation (16) with respect to π_g equal to zero and solving for π_g yields

$$\pi_g^{(k+1)}=n_g^{(k)}\Big/n,$$

where $n_g^{(k)}=\sum_{i=1}^{n}\tau_{ig}^{(k)}$.

4.2.2 Parameters Related to Y

Maximizing (15) with respect to $\boldsymbol{\beta}$ (and possibly to $\boldsymbol{\lambda}$) is equivalent to independently maximizing each of the G expressions

$$\sum_{i=1}^{n}\tau_{ig}^{(k)}\ln q\left(y_i|\boldsymbol{x}_i;\boldsymbol{\beta}_g,\lambda_g\right). \tag{17}$$

The maximization of (17) is equivalent to the maximization problem of the generalized linear models (for the complete data), with the only difference being that each observation $(\boldsymbol{x}_i',y_i)'$ contributes to the log-likelihood with a known weight $\tau_{ig}^{(k)}$. Note that this part of $Q(\boldsymbol{\vartheta};\boldsymbol{\vartheta}^{(k)})$ is also shared by the mixtures of generalized linear models in (8).

Maximization of (17), over $\boldsymbol{\beta}_g$ (and possibly λ_g), can be carried out numerically; details can be found in Wedel and De Sarbo (1995) and Wedel and Kamakura (2001, pp. 120–124), where mixtures of generalized linear models are discussed.

4.2.3 Parameters Related to $\boldsymbol{U}$

Maximizing (15) with respect to $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g$, $g=1,\ldots,G$, is equivalent to independently maximizing each of the G expressions

$$\sum_{i=1}^{n}\tau_{ig}^{(k)}\ln\phi\left(\boldsymbol{u}_i;\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g\right).$$

In particular, we obtain

$$\boldsymbol{\mu}_g^{(k+1)}=\frac{1}{n_g^{(k)}}\sum_{i=1}^{n}\tau_{ig}^{(k)}\boldsymbol{u}_i$$

and

$$\boldsymbol{\Sigma}_g^{(k+1)} = \frac{1}{n_g^{(k)}} \sum_{i=1}^{n} \tau_{ig}^{(k)} \left(\boldsymbol{u}_i - \boldsymbol{\mu}_g^{(k+1)}\right) \left(\boldsymbol{u}_i - \boldsymbol{\mu}_g^{(k+1)}\right)'.$$

4.2.4 Parameters Related to $\boldsymbol{V}$

Maximizing $Q(\boldsymbol{\vartheta}; \boldsymbol{\vartheta}^{(k)})$ over $\boldsymbol{\alpha}_g$, $g = 1, \ldots, G$, subject to the constraints on these parameters, is equivalent to independently maximizing each of the G expressions

$$\sum_{i=1}^{n} \tau_{ig}^{(k)} \ln p\left(\boldsymbol{v}_i; \boldsymbol{\alpha}_g\right) = \sum_{r=1}^{q} \sum_{i=1}^{n} \tau_{ig}^{(k)} \sum_{s=1}^{c_r} v^{rs} \ln \alpha_{grs}.$$

In turn, given the local independence assumption among finite discrete covariates, the maximization of this function with respect to $\boldsymbol{\alpha}_{gr}$, $g = 1, \ldots, G$ and $r = 1, \ldots, q$, subject to the constraints on these parameters, is equivalent to independently maximizing each of the q expressions

$$\sum_{i=1}^{n} \tau_{ig}^{(k)} \sum_{s=1}^{c_r} v_i^{rs} \ln \alpha_{grs} - \eta \left(\sum_{s=1}^{c_r} \alpha_{grs} - 1 \right), \tag{18}$$

where η is a Lagrangian multiplier. Setting the derivative of equation (18) with respect to $\boldsymbol{\alpha}_{gr}$ equal to zero and solving for $\boldsymbol{\alpha}_{gr}$ yields

$$\boldsymbol{\alpha}_{gr}^{(k+1)} = \sum_{i=1}^{n} \tau_{ig}^{(k)} v_i^{rs} \Big/ n_g^{(k)}.$$

4.3 Computational Aspects

The EM algorithm described above is implemented in the **flexCWM** package (Mazza, Punzo and Ingrassia 2013). As far as the response variable is concerned, the present version of the code considers the following four distributions: Gaussian, gamma, Poisson, and binomial. Note that, for the gamma distribution, the noncanonical "log" link is considered to allow the domain of the link function to be the same as the permitted range of the mean of the gamma distribution.

EM initialization. Before running the EM algorithm, the choice of the starting values of the EM algorithm is an important issue. Among the possible initialization strategies (see, e.g., Biernacki, Celeux and Govaert 2000,

Karlis and Xekalaki 2003, and Bagnato and Punzo 2013 for details) a random initialization of $\boldsymbol{\tau}_i^{(0)} = (\tau_{i1}^{(0)}, \ldots, \tau_{iG}^{(0)})'$, $i = 1, \ldots, n$, is repeated t times and the value maximizing the observed-data log-likelihood among these t runs is selected. In each run, each of the n vectors $\boldsymbol{\tau}_i^{(0)}$ can be randomly generated in a "soft" way, by generating G positive values summing to one, or in a "hard" way, by randomly drawn a single observation from a multinomial distribution with probabilities $(1/G, \ldots, 1/G)'$.

EM convergence criterion. The Aitken acceleration (Aitken 1926) is used to estimate the asymptotic maximum of the log-likelihood at each iteration of the EM algorithm. Based on this estimate, a decision can be made regarding whether or not the algorithm has reached convergence, i.e., is the log-likelihood sufficiently close to its estimated asymptotic value (see, e.g., McNicholas, Murphy, McDaid and Frost 2010 for details and implementation).

5. A Simulation Study for Parameter Recovery and Classification Assessment

Following the schemes adopted by Hwang *et al.* (2010) and Punzo (2013), Monte Carlo simulations were conducted to investigate parameter recovery of the EM algorithm. For brevity's sake, attention will only be focused on the component regression coefficients $\boldsymbol{\beta}$.

5.1 Simulation Design

The experimental conditions we considered in the simulation study were the distribution of Y in each group (binomial and Gaussian), sample size ($n = 200$ and $n = 500$), degree of overlap between continuous covariates (see below for details), number of finite discrete covariates ($q = 1$ and $q = 2$), and number of mixture components ($G = 2$ and $G = 3$). The number of continuous covariates was fixed at $p = 2$.

We generated 500 samples, from a generalized linear mixed CWM, at each level of experimental conditions. Subsequently, we fitted all 16,000 samples (2 response component distributions $\times$ 2 sample sizes $\times$ 2 values of overlap between continuous covariates $\times$ 2 numbers of finite discrete covariates $\times$ 2 numbers of mixture components $\times$ 500 replications) with a generalized linear mixed CWM with $G = 2$. The parameters of the generating model were defined as follows.

Mixture weights. Mixture weights were fixed to $\pi_1 = 0.35$ and $\pi_2 = 0.65$ when $G = 2$, and to $\pi_1 = 0.18$, $\pi_2 = 0.32$, and $\pi_3 = 0.50$ when $G = 3$.

Continuous covariates. For the (bivariate) Gaussian distributions of $\boldsymbol{U}|\Omega_g$, $g = 1, \ldots, G$, the covariance matrices were chosen to be identity matrices and the means were selected to lie on the bisector, with $\boldsymbol{\mu}_1 = (0, 0)'$. The mean $\boldsymbol{\mu}_2$ was computed, via a numerical procedure, to guarantee a fixed overlap, with respect to the continuous covariates, of group 1 with group 2. In line with Bagnato, Greselin and Punzo (2014) and Greselin and Punzo (2013), we adopted

$$B = \exp\left(-B^*\right)$$

as a normalized measure of overlap, where

$$B^* = \frac{1}{8}\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)' \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right) + \frac{1}{2} \ln\left(\frac{|\boldsymbol{\Sigma}|}{\sqrt{|\boldsymbol{\Sigma}_1| + |\boldsymbol{\Sigma}_2|}}\right)$$

is the (positive) measure of overlap of Bhattacharyya (1943), with $\boldsymbol{\Sigma} = \left(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2\right)/2$. We remark that B takes values between 0 (absence of overlap) and 1 (complete overlap). In our experiments, we considered two scenarios: $B = 0.05$ and $B = 0.35$. This means $\boldsymbol{\mu}_2 = (3.46164, 3.46164)'$ when $B = 0.05$, and $\boldsymbol{\mu}_2 = (2.04922, 2.04922)'$ when $B = 0.35$. Furthermore, we set $\boldsymbol{\mu}_3 = 2\boldsymbol{\mu}_2$ when $G = 3$ so that the overlap between groups 2 and 3 was equal to the overlap between groups 1 and 2.

Finite discrete numerical covariates. We considered $c_r = 10$ values for each finite discrete numerical covariate, $r = 1, 2$. In each group g, the probabilities $\boldsymbol{\alpha}_{gr}$ were defined according to a binomial distribution on the support $\{0, 1, \ldots, 9\}$: for $G = 2$, the probability of this binomial was fixed to 0.2 for group 1 and to 0.8 for group 2; for $G = 3$, the probabilities were 0.2, 0.5, and 0.8 for groups 1, 2, and 3, respectively.

Response variable. With regard to the local regressions, the intercepts were fixed to $\beta_{0g} = 0$ and the $p + q$ slopes to $\boldsymbol{\beta}_{1g} = (0.1, \ldots, 0.1)'$, $g = 1, \ldots, G$, for both the Gaussian and the binomial cases. The conditional standard deviation σ_g was set to 0.1 in each group in the Gaussian case, while the maximum value was fixed to $M = 20$ in the binomial case. Thus, from a clustering point of view, the local regression models cannot distinguish the clusters because they have the same parameters in all groups. This is a "strong" version of the configuration of assignment dependence discussed in Hennig (2000).

Figure 1 shows an example of a generated data set related to the following combination of experimental conditions: $n = 200$, $G = 3$, $q = 2$, $B = 0.35$,

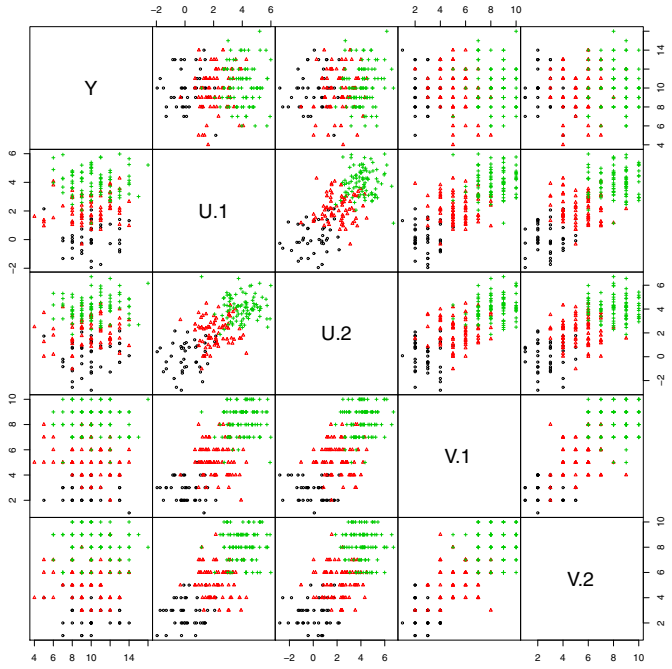


Figure 1. Example of generated data in the scenario considering a binomial response in each group ($n = 200$, $G = 3$, $q = 2$, and $B = 0.35$).

and binomial component response distribution. It is evident here that finding groups in the data is quite difficult; this is an aspect that must be taken into account in the evaluation of the quality of the clustering results.

5.2 Simulation Results

To evaluate the recovery of regression parameter estimates, we computed the mean squared differences of parameters and their estimates as follows:

$$\text{MSD} = \frac{1}{P}\sum_{h=1}^{P}\left(\widehat{\beta}_h - \beta_h\right)^2,$$

where $\widehat{\beta}_h$ and β_h are an estimate and its parameter, respectively, and $P = G\,(1 + p + q)$ is the number of regression parameters.

We performed an analysis of variance (ANOVA) that included MSD as the dependent variable and the five experimental conditions as design factors. For all 16,000 samples, convergence towards the true model was attained and the value of MSD was not substantial. Table 1 presents the

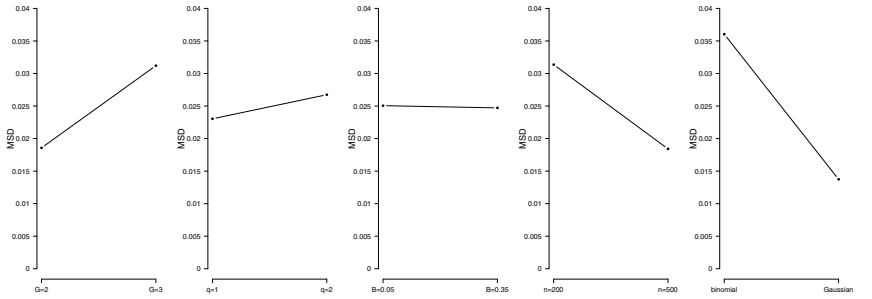
Table 1. ANOVA results for the mean squared differences of regression parameter estimates.

Source	d.f.	Sum of Squares	Mean Square	F value	p-value
# of mixture components (G)	1	0.63778	0.63778	1199.08351	0.00000
# of finite discrete covariates (q)	1	0.05485	0.05485	103.11653	0.00000
overlap on $\boldsymbol{U}$ (B)	1	0.00047	0.00047	0.87508	0.34957
sample size (n)	1	0.67109	0.67109	1261.71648	0.00000
distribution of $Y\|\boldsymbol{x}, \Omega_g$ (D)	1	1.98997	1.98997	3741.34600	0.00000
$G \times q$	1	0.00188	0.00188	3.54397	0.05978
$G \times B$	1	0.00042	0.00042	0.78819	0.37466
$q \times B$	1	0.00111	0.00111	2.09531	0.14777
$G \times n$	1	0.11580	0.11580	217.71417	0.00000
$q \times n$	1	0.00622	0.00622	11.68813	0.00063
$B \times n$	1	0.00027	0.00027	0.50309	0.47816
$G \times D$	1	0.32951	0.32951	619.51938	0.00000
$q \times D$	1	0.00268	0.00268	5.03664	0.02483
$B \times D$	1	0.00001	0.00001	0.01815	0.89283
$n \times D$	1	0.35774	0.35774	672.59070	0.00000
$G \times q \times B$	1	0.00327	0.00327	6.15688	0.01310
$G \times q \times n$	1	0.00009	0.00009	0.17333	0.67718
$G \times B \times n$	1	0.00031	0.00031	0.57464	0.44843
$q \times B \times n$	1	0.00024	0.00024	0.45138	0.50169
$G \times q \times D$	1	0.00243	0.00243	4.57593	0.03244
$G \times B \times D$	1	0.00045	0.00045	0.84651	0.35755
$q \times B \times D$	1	0.00064	0.00064	1.19854	0.27363
$G \times n \times D$	1	0.05860	0.05860	110.17918	0.00000
$q \times n \times D$	1	0.00311	0.00311	5.84089	0.01567
$B \times n \times D$	1	0.00042	0.00042	0.79806	0.37169
$G \times q \times B \times n$	1	0.00260	0.00260	4.87974	0.02719
$G \times q \times B \times D$	1	0.00229	0.00229	4.30485	0.03802
$G \times q \times n \times D$	1	0.00020	0.00020	0.38500	0.53495
$G \times B \times n \times D$	1	0.00042	0.00042	0.78729	0.37493
$q \times B \times n \times D$	1	0.00027	0.00027	0.51239	0.47412
$G \times q \times B \times n \times D$	1	0.00221	0.00221	4.14636	0.04174
Residuals	15,968	8.49317	0.00053		

results of the ANOVA. Except for the overlap, the other main effects of the design factors were statistically significant. In particular, as we can see from Figure 2, the average MSD values improved (decreased) when G decreased (Figure 2(a)), q decreased (Figure 2(b)), n increased (Figure 2(d)), and under component Gaussian responses (Figure 2(e)). Moreover, a number of interaction effects of the design factors were also statistically significant.

5.3 Classification Assessment

The classification performance was also investigated. Table 2 reports the average of the misclassification rates, under all considered experimental conditions, in comparison with model (8) and model (9). Estimates for the latter two models were obtained using the R-package **flexmix** (Grün and Leisch 2008b). The CWM results were always the best while results from mixtures of generalized linear models were the worst. Finally, as expected, the misclassification rates for the CWM generally improved when the over-



(a) G (b) q (c) B (d) n (e) D

Figure 2. Average MSD values across the two levels of each main effect.

lap B (on the continuous covariates) decreased and when the sample size n increased (all other experimental conditions being fixed).

6. A Comparison Among Model Selection Criteria

So far, the number of mixture components G has been treated as known. However, it must be selected in many practical applications. To this end, several information criteria have been proposed in the literature (see, e.g., Fonseca and Cardoso 2005 and Fonseca 2008; 2010); Table 3 lists some of the most commonly used. In this table, $l(\widehat{\boldsymbol{\vartheta}})$ is the (maximized) observed-data log-likelihood, m is the number of free parameters, and MAP $(\widehat{z}_{ig})$ is the *maximum a posteriori* probability operator – assuming a value of 1 if $\max_{j=1,\ldots,G}\{\widehat{z}_{ij}\}$ occurs at component g and 0 otherwise – and $\widehat{\boldsymbol{\vartheta}}$ and $\widehat{z}_{ig}$ are the estimates for $\boldsymbol{\vartheta}$ and z_{ig}, respectively, obtained with the EM algorithm.

Here, we compare the performance of the information criteria in Table 3 in selecting the correct number of latent groups for our model (6). The simulation study is carried out along the same lines of the previous section. In particular, we consider response variables $Y|\boldsymbol{x}, \Omega_g$ having either a Gaussian distribution (results in Table 4 and Table 5) or a binomial distribution (results in Table 6 and Table 7), and we consider 200 replications. Results are reported as average values across replications.

Concerning the Gaussian scenario, Table 4 and Table 5 summarize the obtained results for cases in which $n = 200$ and $n = 500$, respectively. For each replication, the EM algorithm was run for $G \in \{1, 2, 3, 4\}$. The

Table 2. Misclassification rates for three different model-based clustering approaches.Values refer to the averages across 500 replications.

$Y\|\boldsymbol{x},\Omega_g$	n	G	q	B	Mixt. of generalized linear models (8)	Mixt. of generalized linear models with multinomial concomitants (9)	Generalized linear mixed CWM (6)
Gaussian	200	2	1	0.05	0.36355	0.18649	0.00087
				0.35	0.36347	0.18840	0.00694
			2	0.05	0.36214	0.19561	0.00013
				0.35	0.36430	0.19450	0.00080
		3	1	0.05	0.50363	0.32496	0.00691
				0.35	0.49987	0.34182	0.06383
			2	0.05	0.50311	0.32665	0.00406
				0.35	0.50497	0.34517	0.03459
	500	2	1	0.05	0.35928	0.14558	0.00065
				0.35	0.35668	0.15836	0.00647
			2	0.05	0.36010	0.16896	0.00006
				0.35	0.36030	0.17379	0.00064
		3	1	0.05	0.50050	0.31532	0.00598
				0.35	0.50021	0.32740	0.05687
			2	0.05	0.50316	0.33416	0.00364
				0.35	0.50074	0.34410	0.03133
binomial	200	2	1	0.05	0.34396	0.14580	0.00080
				0.35	0.34569	0.15542	0.00668
			2	0.05	0.34935	0.15902	0.00008
				0.35	0.34942	0.15358	0.00079
		3	1	0.05	0.49156	0.27688	0.00660
				0.35	0.49107	0.29979	0.06395
			2	0.05	0.50634	0.33089	0.00377
				0.35	0.50278	0.34818	0.03351
	500	2	1	0.05	0.34539	0.10519	0.00081
				0.35	0.34478	0.12331	0.00668
			2	0.05	0.34636	0.12449	0.00008
				0.35	0.34689	0.13561	0.00061
		3	1	0.05	0.49539	0.26118	0.00616
				0.35	0.49619	0.29164	0.05654
			2	0.05	0.50108	0.32647	0.00356
				0.35	0.50060	0.34313	0.03088

results concern the selection rate, defined here as the proportion of times in which each value of G is selected by the corresponding criterion on the top of the column. The rows related to the true value of G are highlighted in gray; to facilitate performance evaluation, the last row in each table gives the mean selection rate, of each criterion, computed over the true values of G. Regarding Table 4, we first note the very poor performance of the AWE, with a considerable tendency to underestimate the number of groups. Although with an overall lower extent, this tendency is still maintained when the sample size increases to $n = 500$ (see Table 5); in particular, the criterion is unable to see the true number of latent groups when $G = 3$ and $B = 0.35$. On the other hand, the best performing criterion, when $n = 200$, is the AIC_3 (see Table 4). In the case when $n = 500$, all criteria except the AWE and the

Table 3. Definition and key reference for the adopted likelihood-based information criteria.

Information Criterion	Definition	Reference
AIC	$2l(\widehat{\boldsymbol{\vartheta}}) - 2m$	Akaike (1973)
AIC_3	$2l(\widehat{\boldsymbol{\vartheta}}) - 3m$	Bozdogan (1994)
AICc	$\text{AIC} - 2\frac{m(m+1)}{n-m-1}$	Hurvich and Tsai (1989)
AICu	$\text{AICc} - n\ln\frac{n}{n-m-1}$	McQuarrie *et al.* (1997)
AWE	$2l(\widehat{\boldsymbol{\vartheta}}) - 2m\left(\frac{3}{2} + \ln n\right)$	Banfield and Raftery (1993)
BIC	$2l(\widehat{\boldsymbol{\vartheta}}) - m\ln n$	Schwarz (1978)
CAIC	$2l(\widehat{\boldsymbol{\vartheta}}) - m(1 + \ln n)$	Bozdogan (1987)
ICL	$\text{BIC} + \sum_{i=1}^{n}\sum_{g=1}^{G}\text{MAP}(\widehat{z}_{ig})\ln\widehat{z}_{ig}$	Biernacki *et al.* (2000)

AIC performed satisfactorily (see Table 5). Obviously, the performance of the criteria roughly improved for increasing sample size n and decreasing overlap B.

Results concerning the binomial scenario are summarized in Table 6 and Table 7 for sample sizes of $n = 200$ and $n = 500$, respectively. Again in this case, the AWE is the poorest performer.

7. Case Study: The Energy Efficiency Data Set

In this section, we present the results of an application of the generalized linear mixed CWM in modeling the Energy Efficiency data set available at https://archive.ics.uci.edu/ml/datasets/Energy+efficiency. The data set was introduced by Tsanas and Xifara (2012) to study the effect of eight covariates (*relative compactness*, *surface area*, *wall area*, *roof area*, *overall height of the building*, *orientation*, *glazing area*, *glazing area distribution*) on two response variables (*heating load*, *cooling load*) on $n = 768$ residential buildings.

For our purposes, we assumed that data were unlabeled with respect to the dichotomous variable *overall height*, considered here as the group variable. Our interest was to analyze the performance of our model (6) in comparison with model (8) and model (9).

Because there are two pairs of variables (*heating load*, *cooling load*) and (*relative compactness*, *surface area*) that are strongly correlated (with

Table 4. Comparison among model selection criteria: selection rates over 200 replications for $G \in \{1, 2, 3, 4\}$. Response: Gaussian; sample size: $n = 200$.

G	q	B	fitted G	AIC	AICc	AICu	AIC_3	AWE	BIC	CAIC	ICL
2	1	0.05	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
			2	0.810	0.995	1.000	0.990	1.000	1.000	1.000	1.000
			3	0.130	0.005	0.000	0.010	0.000	0.000	0.000	0.000
			4	0.060	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		0.35	1	0.000	0.000	0.000	0.000	0.980	0.000	0.000	0.000
			2	0.805	1.000	1.000	0.995	0.020	1.000	1.000	1.000
			3	0.140	0.000	0.000	0.005	0.000	0.000	0.000	0.000
			4	0.055	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	2	0.05	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
			2	0.940	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			3	0.050	0.000	0.000	0.000	0.000	0.000	0.000	0.000
			4	0.010	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		0.35	1	0.000	0.000	0.000	0.000	0.010	0.000	0.000	0.000
			2	0.960	1.000	1.000	1.000	0.990	1.000	1.000	1.000
			3	0.035	0.000	0.000	0.000	0.000	0.000	0.000	0.000
			4	0.005	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3	1	0.05	1	0.000	0.000	0.000	0.000	0.490	0.000	0.000	0.000
			2	0.000	0.025	0.065	0.005	0.510	0.135	0.355	0.125
			3	0.680	0.955	0.930	0.925	0.000	0.860	0.640	0.870
			4	0.320	0.020	0.005	0.070	0.000	0.005	0.005	0.005
		0.35	1	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000
			2	0.025	0.385	0.930	0.245	0.000	0.975	1.000	0.975
			3	0.710	0.615	0.070	0.740	0.000	0.025	0.000	0.025
			4	0.265	0.000	0.000	0.015	0.000	0.000	0.000	0.000
	2	0.05	1	0.000	0.000	0.000	0.000	0.075	0.000	0.000	0.000
			2	0.005	0.060	0.170	0.005	0.925	0.070	0.170	0.065
			3	0.865	0.940	0.830	0.940	0.000	0.930	0.830	0.935
			4	0.130	0.000	0.000	0.055	0.000	0.000	0.000	0.000
		0.35	1	0.000	0.000	0.000	0.000	0.955	0.000	0.000	0.000
			2	0.000	0.625	0.990	0.075	0.045	0.805	0.990	0.810
			3	0.810	0.375	0.010	0.890	0.000	0.195	0.010	0.190
			4	0.190	0.000	0.000	0.035	0.000	0.000	0.000	0.000
means over gray rows				**0.823**	**0.860**	**0.730**	**0.935**	**0.376**	**0.751**	**0.685**	**0.753**

Table 5. Comparison among model selection criteria: selection rates over 200 replications for $G \in \{1, 2, 3, 4\}$. Response: Gaussian; sample size: $n = 500$.

G	q	B	fitted G	AIC	AICc	AICu	AIC_3	AWE	BIC	CAIC	ICL
2	1	0.05	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
			2	0.770	0.965	1.000	0.995	1.000	1.000	1.000	1.000
			3	0.150	0.035	0.000	0.005	0.000	0.000	0.000	0.000
			4	0.080	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		0.35	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
			2	0.770	0.945	1.000	0.995	1.000	1.000	1.000	1.000
			3	0.140	0.050	0.000	0.005	0.000	0.000	0.000	0.000
			4	0.090	0.005	0.000	0.000	0.000	0.000	0.000	0.000
	2	0.05	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
			2	0.915	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			3	0.065	0.000	0.000	0.000	0.000	0.000	0.000	0.000
			4	0.020	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		0.35	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
			2	0.955	0.990	1.000	0.995	1.000	1.000	1.000	1.000
			3	0.030	0.010	0.000	0.005	0.000	0.000	0.000	0.000
			4	0.015	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3	1	0.05	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
			2	0.000	0.000	0.000	0.000	0.345	0.000	0.000	0.000
			3	0.770	0.960	0.985	0.970	0.655	1.000	1.000	1.000
			4	0.230	0.040	0.015	0.030	0.000	0.000	0.000	0.000
		0.35	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
			2	0.000	0.000	0.000	0.000	1.000	0.085	0.295	0.175
			3	0.805	0.980	0.995	0.995	0.000	0.915	0.705	0.825
			4	0.195	0.020	0.005	0.005	0.000	0.000	0.000	0.000
	2	0.05	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
			2	0.000	0.000	0.000	0.000	0.180	0.000	0.000	0.000
			3	0.865	0.980	0.980	0.980	0.820	0.990	0.990	0.990
			4	0.135	0.020	0.020	0.020	0.000	0.010	0.010	0.010
		0.35	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
			2	0.000	0.000	0.000	0.000	1.000	0.015	0.015	0.015
			3	0.880	0.985	0.985	0.985	0.000	0.985	0.985	0.985
			4	0.120	0.015	0.015	0.015	0.000	0.000	0.000	0.000
means over gray rows				**0.841**	**0.976**	**0.993**	**0.989**	**0.684**	**0.986**	**0.960**	**0.975**

Table 6. Comparison among model selection criteria: selection rates over 200 replications for $G \in \{1, 2, 3, 4\}$. Response: binomial; sample size: $n = 200$.

G	q	B	fitted G	AIC	AICc	AICu	AIC_3	AWE	BIC	CAIC	ICL
2	1	0.05	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
			2	0.980	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			3	0.020	0.000	0.000	0.000	0.000	0.000	0.000	0.000
			4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		0.35	1	0.000	0.000	0.000	0.000	0.930	0.000	0.000	0.000
			2	0.940	1.000	1.000	1.000	0.070	1.000	1.000	1.000
			3	0.045	0.000	0.000	0.000	0.000	0.000	0.000	0.000
			4	0.015	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	2	0.05	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
			2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
			4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		0.35	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
			2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
			4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3	1	0.05	1	0.000	0.000	0.000	0.000	0.360	0.000	0.000	0.000
			2	0.000	0.000	0.015	0.000	0.645	0.050	0.185	0.050
			3	0.920	0.990	0.990	0.980	0.000	0.955	0.820	0.955
			4	0.085	0.015	0.000	0.025	0.000	0.000	0.000	0.000
		0.35	1	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000
			2	0.005	0.240	0.860	0.170	0.000	0.950	0.995	0.950
			3	0.845	0.760	0.140	0.820	0.000	0.050	0.005	0.050
			4	0.150	0.000	0.000	0.010	0.000	0.000	0.000	0.000
	2	0.05	1	0.000	0.000	0.000	0.000	0.050	0.000	0.000	0.000
			2	0.000	0.040	0.085	0.000	0.950	0.045	0.140	0.045
			3	0.920	0.960	0.915	0.960	0.000	0.955	0.860	0.955
			4	0.080	0.000	0.000	0.040	0.000	0.000	0.000	0.000
		0.35	1	0.000	0.000	0.000	0.000	0.850	0.000	0.000	0.000
			2	0.000	0.490	0.950	0.085	0.150	0.760	0.975	0.765
			3	0.870	0.510	0.050	0.890	0.000	0.240	0.025	0.235
			4	0.130	0.000	0.000	0.025	0.000	0.000	0.000	0.000
mean over the gray rows				**0.934**	**0.903**	**0.762**	**0.956**	**0.384**	**0.775**	**0.714**	**0.774**

Table 7. Comparison among model selection criteria: selection rates over 200 replications for $G \in \{1, 2, 3, 4\}$. Response: binomial; sample size: $n = 500$.

G	q	B	fitted G	AIC	AICc	AICu	AIC_3	AWE	BIC	CAIC	ICL
2	1	0.05	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
			2	0.980	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			3	0.015	0.000	0.000	0.000	0.000	0.000	0.000	0.000
			4	0.005	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		0.35	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
			2	0.935	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			3	0.055	0.000	0.000	0.000	0.000	0.000	0.000	0.000
			4	0.010	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	2	0.05	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
			2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
			4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		0.35	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
			2	0.995	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			3	0.005	0.000	0.000	0.000	0.000	0.000	0.000	0.000
			4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3	1	0.05	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
			2	0.000	0.000	0.000	0.000	0.180	0.000	0.000	0.000
			3	0.880	0.970	0.980	0.975	0.820	0.995	1.000	1.000
			4	0.120	0.030	0.020	0.025	0.000	0.005	0.000	0.000
		0.35	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
			2	0.000	0.000	0.000	0.000	1.000	0.035	0.245	0.115
			3	0.920	0.995	0.995	0.995	0.000	0.965	0.755	0.885
			4	0.080	0.005	0.005	0.005	0.000	0.000	0.000	0.000
	2	0.05	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
			2	0.000	0.000	0.000	0.000	0.125	0.000	0.000	0.000
			3	0.940	0.965	0.980	0.965	0.875	0.985	0.985	0.985
			4	0.060	0.035	0.020	0.035	0.000	0.015	0.015	0.015
		0.35	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
			2	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000
			3	0.955	0.995	0.995	0.995	0.000	1.000	1.000	1.000
			4	0.045	0.005	0.005	0.005	0.000	0.000	0.000	0.000
mean over the gray rows				**0.951**	**0.991**	**0.994**	**0.991**	**0.712**	**0.993**	**0.968**	**0.984**

Table 8. Description of the variables in the Energy Efficiency data set.

Variable description	Original description	Present notation	Number of values (if finite discrete)
Heating Load	Response	Y	
Cooling Load	Response	Not used	
Relative Compactness	Covariate	U	
Surface Area	Covariate	Not used	
Wall Area	Covariate	V_1	7
Roof Area	Covariate	V_2	4
Orientation	Covariate	V_3	4
Glazing Area	Covariate	V_4	4
Glazing Area Distribution	Covariate	V_5	6
Overall Height	Covariate	Group variable	2

Table 9. Energy Efficiency data set. Confusion matrices for three mixture-based approaches.

(a) Linear Gaussian CWM

True \ Est.	1	2
Low Height	384	–
High Height	–	384

(b) Mixture of linear (Gaussian) models

True \ Est.	1	2
Low Height	350	34
High Height	191	193

(c) Mixture of linear models with multinomial concomitants

True \ Est.	1	2
Low Height	384	–
High Height	64	320

correlation coefficient of 0.976 and -0.992, respectively), we reduced the number of considered variables by only taking into account the variables *heating load* and *relative compactness*. Table 8 provides information about the original variables and the variables taken into account in our case study (according to the notation introduced in Section 2); for each finite discrete variable, the number of values is also given.

Table 9 shows the confusion matrices for all of the fitted models. They were directly estimated with $G = 2$ and using a Gaussian response variable in each group. As we can see from Table 9, our model produced a perfect classification while the other competitive approaches did not work well.

8. Concluding Remarks

In this paper, we have introduced the generalized linear mixed cluster-weighted model. Different from previous work about cluster-weighted models, it allows modelization of various types of response variables as well as covariates of mixed-type (finite discrete and continuous). Furthermore, we have described an EM algorithm for parameter estimation and have given sufficient conditions for model identifiability under the assumption of Gaus-

sian covariates. Broad simulation studies have been presented in order to investigate the parameter recovery of the algorithm and the performance of different model selection criteria in selecting the number of mixture components. When the proposed model was applied to real data in Section 7, it demonstrated optimal classification performance in contrast to other existing mixture-based approaches.

Appendix

Proof of Theorem 1. The proof builds upon results given in Hennig (2000). Consider the class of models defined in (10) and prove that the equality

$$\sum_{g=1}^{G} q(y|\boldsymbol{u};\boldsymbol{\beta}_g,\lambda_g)\phi(\boldsymbol{u};\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g)\pi_g = \sum_{s=1}^{\widetilde{G}} q(y|\boldsymbol{u};\widetilde{\boldsymbol{\beta}}_s,\widetilde{\lambda}_s)\phi(\boldsymbol{u};\widetilde{\boldsymbol{\mu}}_s,\widetilde{\boldsymbol{\Sigma}}_s)\widetilde{\pi}_s \tag{19}$$

holds for almost all $\boldsymbol{u} \in \mathbb{R}^p$ and for all $y \in \mathcal{Y}$ if and only if $G = \widetilde{G}$ and for each $g \in \{1,\ldots,G\}$ there exists $s \in \{1,\ldots,G\}$ such that $\boldsymbol{\beta}_g = \widetilde{\boldsymbol{\beta}}_s$, $\lambda_g = \widetilde{\lambda}_s$, $\boldsymbol{\mu}_g = \widetilde{\boldsymbol{\mu}}_s$, $\boldsymbol{\Sigma}_g = \widetilde{\boldsymbol{\Sigma}}_s$ and $\pi_g = \widetilde{\pi}_s$.

Summing (or integrating) each side of the equality (19) over $\mathcal{Y}$ yields

$$\sum_{g=1}^{G} \phi(\boldsymbol{u};\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g)\pi_g = \sum_{s=1}^{\widetilde{G}} \phi(\boldsymbol{u};\widetilde{\boldsymbol{\mu}}_s,\widetilde{\boldsymbol{\Sigma}}_s)\widetilde{\pi}_s. \tag{20}$$

Let us set

$$p(\boldsymbol{u};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\pi}) = \sum_{g=1}^{G} \phi(\boldsymbol{u};\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g)\pi_g$$

and

$$p(\boldsymbol{u};\widetilde{\boldsymbol{\mu}},\widetilde{\boldsymbol{\Sigma}},\widetilde{\boldsymbol{\pi}}) = \sum_{s=1}^{\widetilde{G}} \phi(\boldsymbol{u};\widetilde{\boldsymbol{\mu}}_s,\widetilde{\boldsymbol{\Sigma}}_s)\widetilde{\pi}_s,$$

where $\boldsymbol{\mu} = \{\boldsymbol{\mu}_g;\ g = 1,\ldots,G\}$, $\boldsymbol{\Sigma} = \{\boldsymbol{\Sigma}_g;\ g = 1,\ldots,G\}$ and $\boldsymbol{\pi} = \{\pi_g,;\ g = 1,\ldots,G\}$; analogous notation applies for $\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}$ and $\widetilde{\boldsymbol{\pi}}$. Afterwards, based on Bayes' theorem, we get

$$\begin{aligned} p(\Omega_g|\boldsymbol{u};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\pi}) &= \frac{\phi(\boldsymbol{u};\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g)\pi_g}{\sum_{j=1}^{G}\phi(\boldsymbol{u};\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)\pi_j} \\ &= \frac{\phi(\boldsymbol{u};\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g)\pi_g}{p(\boldsymbol{u};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\pi})},\ g = 1,\ldots,G, \end{aligned} \tag{21}$$

and thus rewrite model (7) as

$$p(\boldsymbol{u}, y; \boldsymbol{\vartheta}) = p(\boldsymbol{u}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) \sum_{g=1}^{G} q(y|\boldsymbol{u}; \boldsymbol{\beta}_g, \lambda_g) p(\Omega_g|\boldsymbol{u}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})$$
$$= p(\boldsymbol{u}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) p(y|\boldsymbol{u}; \boldsymbol{\vartheta}) \tag{22}$$

where

$$p(y|\boldsymbol{u}; \boldsymbol{\vartheta}) = \sum_{g=1}^{G} q(y|\boldsymbol{u}; \boldsymbol{\beta}_g, \lambda_g) p(\Omega_g|\boldsymbol{u}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}), \quad y \in \mathcal{Y}. \tag{23}$$

Now, the class of models defined by (23) is identifiable, for almost all $\boldsymbol{u} \in \mathbb{R}^p$, if the equality

$$\sum_{g=1}^{G} q(y|\boldsymbol{u}; \boldsymbol{\beta}_g, \lambda_g) p(\Omega_g|\boldsymbol{u}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{s=1}^{\widetilde{G}} q(y|\boldsymbol{u}; \widetilde{\boldsymbol{\beta}}_s, \widetilde{\lambda}_s) p(\Omega_s|\boldsymbol{u}; \widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}, \widetilde{\boldsymbol{\pi}})$$

implies that $G = \widetilde{G}$ and for each $g \in \{1, \ldots, G\}$ there exists $s \in \{1, \ldots, G\}$ such that $\boldsymbol{\beta}_g = \widetilde{\boldsymbol{\beta}}_s$, $\lambda_g = \widetilde{\lambda}_s$, $\boldsymbol{\mu}_g = \widetilde{\boldsymbol{\mu}}_s$, $\boldsymbol{\Sigma}_g = \widetilde{\boldsymbol{\Sigma}}_s$ and $\pi_g = \widetilde{\pi}_s$.

In Section 2.1, we remarked that the expected value μ_g of $Y|\Omega_g$ is related to the covariates $\boldsymbol{X}$ through the relation $h(\mu_g) = \beta_{0g} + \boldsymbol{\beta}'_{1g}\boldsymbol{x}$, $g = 1, \ldots, G$. Let us introduce the set

$$\mathcal{U} = \Big\{ \boldsymbol{u} \in \mathbb{R}^p : \text{ for each } g, j \in \{1, \ldots, G\}, \text{ and } s, t \in \{1, \ldots, \widetilde{G}\} :$$
$$\beta_{0g} + \boldsymbol{\beta}'_{1g}\boldsymbol{u} = \beta_{0j} + \boldsymbol{\beta}'_{1j}\boldsymbol{u} \Rightarrow (\beta_{0g}, \boldsymbol{\beta}'_{1g})' = (\beta_{0j}, \boldsymbol{\beta}'_{1j})',$$
$$\beta_{0g} + \boldsymbol{\beta}'_{1g}\boldsymbol{u} = \widetilde{\beta}_{0s} + \widetilde{\boldsymbol{\beta}}'_{1s}\boldsymbol{u} \Rightarrow (\beta_{0g}, \boldsymbol{\beta}'_{1g})' = (\widetilde{\beta}_{0s}, \widetilde{\boldsymbol{\beta}}'_{1s})',$$
$$\widetilde{\beta}_{0s} + \widetilde{\boldsymbol{\beta}}'_{1s}\boldsymbol{u} = \widetilde{\beta}_{0t} + \widetilde{\boldsymbol{\beta}}'_{1t}\boldsymbol{u} \Rightarrow (\widetilde{\beta}_{0s}, \widetilde{\boldsymbol{\beta}}'_{1s})' = (\widetilde{\beta}_{0t}, \widetilde{\boldsymbol{\beta}}'_{1t})' \Big\}.$$

Since $(\boldsymbol{\beta}'_g, \lambda_g)' \neq (\boldsymbol{\beta}'_j, \lambda_j)'$ for $g \neq j$ according to (10), it follows that the quantities $(\beta_{0g} + \boldsymbol{\beta}'_{1g}\boldsymbol{u}, \lambda_g)$, $g = 1, \ldots, G$ are pairwise distinct for all $\boldsymbol{u} \in \mathcal{U}$ and the set $\mathcal{U}$ has probability one according to the p-variate Gaussian distribution (indeed, the complement of $\mathcal{U}$, i.e. $\mathbb{R}^p \setminus \mathcal{U}$, is formed by a finite set of hyperplanes of $\mathbb{R}^p$ and thus $\mathbb{R}^p \setminus \mathcal{U}$ has null measure).

For any fixed $\boldsymbol{u} \in \mathcal{U}$, according to (21), $\{p(\Omega_1|\boldsymbol{u}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}), \ldots, p(\Omega_G|\boldsymbol{u}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})\}$ and $\{p(\Omega_1|\boldsymbol{u}; \widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}, \widetilde{\boldsymbol{\pi}}), \ldots, p(\Omega_{\widetilde{G}}|\boldsymbol{u}; \widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}, \widetilde{\boldsymbol{\pi}})\}$ are sets of positive numbers summing to one. It follows that, for each $\boldsymbol{u} \in \mathcal{U}$, the density $p(y|\boldsymbol{u}; \boldsymbol{\vartheta})$ given in (23) is a mixture of distributions of kind (11);

then it is identifiable due to the assumptions of the theorem. Thus $G = \widetilde{G}$ and there exists $s \in \{1, \ldots, G\}$ such that

$$\boldsymbol{\beta}_g = \widetilde{\boldsymbol{\beta}}_s, \quad \lambda_g = \widetilde{\lambda}_s \quad \text{and} \quad p(\Omega_g|\boldsymbol{u}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = p(\Omega_s|\boldsymbol{u}; \widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}, \widetilde{\boldsymbol{\pi}}). \quad (24)$$

Moreover, since $p(\Omega_g|\boldsymbol{u}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})$ and $p(\Omega_s|\boldsymbol{u}; \widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}, \widetilde{\boldsymbol{\pi}})$ are defined according to (21), from (24) and (20) we get:

$$\begin{aligned}
\pi_g &= \int_{\mathcal{U}} \pi_g \phi(\boldsymbol{u}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) d\boldsymbol{u} \\
&= \int_{\mathcal{U}} \frac{\pi_g \phi(\boldsymbol{u}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)}{\sum_{j=1}^{G} \phi(\boldsymbol{u}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\pi_j} \left(\sum_{j=1}^{G} \phi(\boldsymbol{u}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\pi_j \right) d\boldsymbol{u} \\
&= \int_{\mathcal{U}} p(\Omega_g|\boldsymbol{u}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) \left(\sum_{t=1}^{\widetilde{G}} \phi(\boldsymbol{u}; \widetilde{\boldsymbol{\mu}}_t, \widetilde{\boldsymbol{\Sigma}}_t)\widetilde{\pi}_t \right) d\boldsymbol{u} \\
&= \int_{\mathcal{U}} p(\Omega_s|\boldsymbol{u}; \widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}, \widetilde{\boldsymbol{\pi}}) \left(\sum_{t=1}^{\widetilde{G}} \phi(\boldsymbol{u}; \widetilde{\boldsymbol{\mu}}_s t, \widetilde{\boldsymbol{\Sigma}}_t)\widetilde{\pi}_t \right) d\boldsymbol{u} \\
&= \int_{\mathcal{U}} \frac{\widetilde{\pi}_s \phi(\boldsymbol{u}; \widetilde{\boldsymbol{\mu}}_s, \widetilde{\boldsymbol{\Sigma}}_s)}{\sum_{t=1}^{\widetilde{G}} \phi(\boldsymbol{u}; \widetilde{\boldsymbol{\mu}}_t, \widetilde{\boldsymbol{\Sigma}}_t)\widetilde{\pi}_t} \left(\sum_{t=1}^{\widetilde{G}} \phi(\boldsymbol{u}; \widetilde{\boldsymbol{\mu}}_t, \widetilde{\boldsymbol{\Sigma}}_t \widetilde{\pi}_t \right) d\boldsymbol{u} \\
&= \int_{\mathcal{U}} \widetilde{\pi}_s \phi(\boldsymbol{u}; \widetilde{\boldsymbol{\mu}}_s, \widetilde{\boldsymbol{\Sigma}}_s) d\boldsymbol{u} = \widetilde{\pi}_s.
\end{aligned}$$

Thus for the same pair (g, s) in (24), it results $\pi_g = \widetilde{\pi}_s$. Finally, we get

$$\begin{aligned}
\phi(\boldsymbol{u}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) &= \frac{p(\Omega_g|\boldsymbol{u}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})}{\pi_g} \left(\sum_{g=1}^{G} \phi(\boldsymbol{u}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)\pi_g \right) \\
&= \frac{p(\Omega_s|\boldsymbol{u}; \widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}, \widetilde{\boldsymbol{\pi}})}{\widetilde{\pi}_s} \left(\sum_{s=1}^{\widetilde{G}} \phi(\boldsymbol{u}; \widetilde{\boldsymbol{\mu}}_s, \widetilde{\boldsymbol{\Sigma}}_s)\widetilde{\pi}_s \right) \\
&= \phi(\boldsymbol{u}; \widetilde{\boldsymbol{\mu}}_s, \widetilde{\boldsymbol{\Sigma}}_s).
\end{aligned}$$

From the identifiability of Gaussian distributions, again for the same pair (g, s) in (24), it follows that

$$\boldsymbol{\mu}_g = \widetilde{\boldsymbol{\mu}}_s \quad \text{and} \quad \boldsymbol{\Sigma}_g = \widetilde{\boldsymbol{\Sigma}}_s.$$

This completes the proof.

■

References

AITKEN, A.C. (1926), "On Bernoulli's Numerical Solution of Algebraic Equations", in *Proceedings of the Royal Society of Edinburgh*, Vol. 46, pp. 289–305.

AKAIKE, H. (1973), "Information Theory and an Extension of Maximum Likelihood Principle", in *Second International Symposium on Information Theory*, eds. B.N. Petrov and F. Csaki, Budapest: Akademiai Kiado, pp. 267–281.

BAGNATO, L., and PUNZO, A. (2013), "Finite Mixtures of Unimodal Beta and Gamma Densities and the k-bumps Algorithm", *Computational Statistics*, *28(4)*, 1571–1597.

BAGNATO, L., GRESELIN, F., and PUNZO, A. (2014), "On the Spectral Decomposition in Normal Discriminant Analysis", *Communications in Statistics - Simulation and Computation*, *43(6)*, 1471–1489.

BANFIELD, J.D., and RAFTERY, A.E. (1993), "Model-based Gaussian and non-Gaussian Clustering", *Biometrics*, *49(3)*, 803–821.

BHATTACHARYYA, A. (1943), "On a Measure of Divergence Between Two Statistical Populations Defined by Their Probability Distributions", *Bulletin of the Calcutta Mathematical Society*, *35(4)*, 99–109.

BIERNACKI, C., CELEUX, G., and GOVAERT, G. (2000), "Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22(7)*, 719–725.

BIERNACKI, C., CELEUX, G., and GOVAERT, G. (2003), "Choosing Starting Values for the EM Algorithm for Getting the Highest Likelihood in Multivariate Gaussian Mixture Models", *Computational Statistics and Data Analysis*, *41(3-4)*, 561–575.

BOZDOGAN, H. (1987), " Model Selection and Akaikes's Information Criterion (AIC): The General Theory and Its Analytical Extensions", *Psychometrika*, *52*, 345–370.

BOZDOGAN, H. (1994), "Theory & Methodology of Time Series Analysis", in *Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach* (Vol. 1), Dordrecht: Kluwer Academic Publishers.

DEMPSTER, A.P., LAIRD, N.M., and RUBIN, D.B. (1977), "Maximum Likelihood From Incomplete Data Via the EM Algorithm", *Journal of the Royal Statistical Society. Series B (Methodological)*, *39(1)*, 1–38.

FOLLMANN, D.A., and LAMBERT, D. (1991), "Identifiability of Finite Mixtures of Logistic Regression Models", *Journal of Statistical Planning and Inference*, *27(3)*, 375–381.

FONSECA, J.R.S. (2008), "The Application of Mixture Modeling and Information Criteria for Discovering Patterns of Coronary Heart Disease", *Journal of Applied Quantitative Methods*, *3(4)*, 292–303.

FONSECA, J.R.S. (2010), "On the Performance of Information Criteria in Latent Segment Models", *World Academy of Science, Engineering and Technology*, *63*, 2010.

FONSECA, J.R.S., and CARDOSO, M.G.M.S. (2005), "Retail Clients Latent Segments", in *Progress in Artificial Intelligence*, Berlin Heidelberg: Springer-Verlag, pp. 348–358.

FRÜHWIRTH-SCHNATTER, S. (2006), *Finite Mixture and Markov Switching Models*, New York: Springer.

GERSHENFELD, N. (1997), "Nonlinear Inference and Cluster-Weighted Modeling", *Annals of the New York Academy of Sciences*, *808(1)*, 18–24.

GERSHENFELD, N. (1999), *The Nature of Mathematical Modelling*, Cambridge: Cambridge University Press.

GERSHENFELD, N., SCHÖNER, B., and METOIS, E. (1999), "Cluster-Weighted Modelling for Time-Series Analysis", *Nature*, *397*, 329–332.

GRESELIN, F., and PUNZO, A. (2013), "Closed Likelihood Ratio Testing Procedures to Assess Similarity of Covariance Matrices", *The American Statistician*, *67(3)*, 117–128.

GRÜN, B., and LEISCH, F. (2008a), "Finite Mixtures of Generalized Linear Regression Models", in *Recent Advances in Linear Models and Related Areas - Essays in Honour of Helge Toutenburg Shalabh*, ed. C. Heumann, Heidelberg: Springer Physica Verlag, pp. 205–230.

GRÜN, B., and LEISCH, F. (2008b), " **FlexMix** Version 2: Finite Mixtures with Concomitant Variables and Varying and Constant Parameters", *Journal of Statistical Software*, *28(4)*, 1–35.

HENNIG, C. (2000), "Identifiablity of Models for Clusterwise Linear Regression", *Journal of Classification*, *17(2)*, 273–296.

HENNIG, C., and LIAO, T.F. (2013), "How to Find an Appropriate Clustering for Mixed Type Variables with Application to Socio-Economic Stratification", *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *62(3)*, 1–25.

HURVICH, C.M., and TSAI, C.L. (1989), "Regression and Time Series Model Selection in Small Samples", *Biometrika*, *76(2)*, 297–307.

HWANG, H., MALHOTRA, N.K., KIM, Y., TOMIUK, M.A., and HONG, S. (2010), "A Comparative Study on Parameter Recovery of Three Approaches to Structural Equation Modeling", *Journal of Marketing Research*, *47(4)*, 699–712.

INGRASSIA, S., MINOTTI, S.C., and VITTADINI, G. (2012), "Local Statistical Modeling Via the Cluster-Weighted Approach with Elliptical Distributions", *Journal of Classification*, *29(3)*, 363–401.

INGRASSIA, S., MINOTTI, S.C., and PUNZO, A. (2014), "Model-Based Clustering Via Linear Cluster-Weighted Models", *Computational Statistics and Data Analysis*, *71*, 159–182.

KARLIS, D., and XEKALAKI, E. (2003), "Choosing Initial Values for the EM Algorithm for Finite Mixtures", *Computational Statistics and Data Analysis*, *41(3–4)*, 577–590.

MAZZA, A., PUNZO, A., and INGRASSIA, S. (2013), ***flexCWM****: Flexible Cluster-Weighted Modeling*, available at http://cran.fhcrc.org/web/packages/flexCWM/index.html.

MCCULLAGH, P., and NELDER, J.A. (2000), *Generalized Linear Models* (2nd ed.), Boca Raton: Chapman and Hall.

MCLACHLAN, G.J. (1997), "On the EM Algorithm for Overdispersed Count Data", *Statistical Methods in Medical Research*, *6(1)*, 76–98.

MCLACHLAN, G.J., and PEEL, D. (2000), *Finite Mixture Models*, New York: John Wiley and Sons.

MCNICHOLAS, P.D., MURPHY, T.B., MCDAID, A.F., and FROST, D. (2010), "Serial and Parallel Implementations of Model-Based Clustering Via Parsimonious Gaussian Mixture Models", *Computational Statistics and Data Analysis*, *54(3)*, 711–723.

MCQUARRIE, A., SHUMWAY, R., and TSAI, C.L. (1997), "The Model Selection Criterion AICu", *Statistics and Probability Letters*, *34(3)*, 285–292.

PUNZO, A. (2014), "Flexible Mixture Modeling with the Polynomial Gaussian Cluster-Weighted Model", *Statistical Modelling*, *14(3)*, 257–291.

R CORE TEAM (2013), R*: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing.

SCHÖNER, B. (2000), "Probabilistic Characterization and Synthesis of Complex Data Driven Systems", Technical Report, Ph.D. Thesis, MIT, Cambridge.

SCHÖNER, B., and GERSHENFELD, N. (2001), "Cluster Weighted Modeling: Probabilistic Time Series Prediction, Characterization, and Synthesism", in *Nonlinear Dynamics and Statistics*, ed. A. Mees, Boston: Birkhauser, pp. 365–385.

SCHWARZ, G. (1978), "Estimating the Dimension of a Model", *The Annals of Statistics*, *6(2)*, 461–464.

SUBEDI, S., PUNZO, A., INGRASSIA, S., and MCNICHOLAS, P.D. (2013), "Clustering and Classification Via Cluster-Weighted Factor Analyzers", *Advances in Data Analysis and Classification*, *7(1)*, 5–40.

TEICHER, H. (1963), "Identifiability of Finite Mixtures", *Annals of Mathematical Statistics*, *34(4)*, 1265–1269.

TITTERINGTON, D.M., SMITH, A.F.M., and MAKOV, U.E. (1985), *Statistical Analysis of Finite Mixture Distributions*, New York: John Wiley and Sons.

TSANAS, A., and XIFARA, A. (2012), "Accurate Quantitative Estimation of Energy Performance of Residential Buildings Using Statistical Machine Learning Tools", *Energy and Buildings*, *49*, 560–567.

VERMUNT, J.K., and MAGIDSON, J. (2002), "Latent Class Cluster Analysis", in *Applied Latent Class Analysis*, eds. J.A. Hagenaars and A.L. McCutcheon, Cambridge: Cambridge University Press, pp. 89–106.

WANG, P. (1994), "Mixed Regression Models for Discrete Data", Technical Report, Ph.D. Thesis, University of British Columbia, Vancouver.

WANG, P., PUTERMAN, M.L., COCKBURN, M.L., and LE, N.D. (1996), "Mixed Poisson Regression Models with Covariate Dependent Rates", *Biometrics*, *52(2)*, 381–400.

WEDEL, M. (2002), "Concomitant Variables in Finite Mixture Models", *Statistica Neerlandica*, *56(3)*, 362–375.

WEDEL, M., and DE SARBO, W. (1995), "A Mixture Likelihood Approach for Generalized Linear Models", *Journal of Classification*, *12(3)*, 21–55.

WEDEL, M., and KAMAKURA, W.A. (2001), *Market Segmentation: Conceptual and Methodological Foundations* (2nd ed.), Boston MA: Kluwer Academic Publishers.